Abstract in Portuguese - Resumo em português

The Portuguese translation of the abstract was done using Google Translate and corrected by Débora Y. C. Brandt, who is a native Portuguese speaker.

A tradução do resumo deste artigo para o português foi feita usando o Google Tradutor, seguida de correção manual por Débora Y. C. Brandt, que tem o português como língua materna.

O grafo de recombinação ancestral (GRA, ou ARG na sigla em inglês) é uma estrutura que descreve o conjunto de genealogias locais ao longo do genoma, para um conjunto de sequências de DNA amostradas. Métodos computacionais desenvolvidos recentemente geraram um progresso impressionante na possibilidade de estimar genealogias de todo o genoma para um grande número de amostras. Além de inferir um único ARG, alguns desses métodos também podem fornecer diversos ARG amostrados de uma distribuição *a posteriori*. Obter uma boa amostra de ARGs é crucial para quantificar a incerteza estatística e para estimar parâmetros populacionais, como tamanho efetivo da população, taxa de mutação e idade de alelos. Neste trabalho, usamos simulações sob o modelo coalescente neutro padrão para comparar as estimativas de tempos de coalescência par-a-par de três programas amplamente utilizados para a inferência de ARGs: ARGweaver, Relate e tsinfer+tsdate. Comparamos 1) os tempos de coalescência simulados com os tempos inferidos em cada *locus*; 2) a distribuição dos tempos de coalescência par-a-par para todos os *loci* com a distribuição exponencial que seria esperada; 3) se os tempos de coalescência amostrados possuem as propriedades esperadas de uma distribuição *a posteriori* bem calibrada. Descobrimos que os tempos de coalescência inferidos locus-a-locus pelo programa ARGweaver são os mais precisos, e que geralmente os tempos de coalescência inferidos pelo programa Relate são mais precisos do que os inferidos por tsinfer+tsdate. No entanto, os três métodos tendem a superestimar tempos de coalescência baixos e subestimar os altos. Por fim, as amostras da distribuição *a posteriori* geradas pelo programa ARGweaver refletem uma distribuição mais próxima da distribuição *a posteriori* esperada do que as amostras geradas pelo programa Relate, mas essa precisão mais alta é acompanhada de custo computacional muito mais elevado. Portanto, a escolha do melhor método a ser usado depende do número e comprimento das sequências amostradas, e do objetivo das análises em que se deseja usar o ARG. Por fim, oferecemos algumas recomendações de uso desses métodos para diferentes fins.

Palavras-chave: Grafo de recombinação ancestral; ARGweaver; Relate; tsinfer; tsdate; simulação; calibração; distribuição *a posteriori*

Evaluating MCMC Convergence

To evaluate MCMC convergence in ARGweaver and Relate, we run these programs five independent times for the same simulated sequence of 5Mb. We do this for each simulation scenario and evaluate convergence by analysing various statistics extracted at each iteration. For ARGweaver, we analyse statistics from in the *.stats* file, described below. Relate does not generate a similar output, so we extract a subset of the pairwise coalescence times at each MCMC iteration to evaluate convergence. We also evaluate convergence based on selected pairwise coalescence times in ARGweaver, for comparison. Using these statistics extracted at each iteration, we evaluate MCMC convergence by analysing 1) trace plots, 2) autocorrelation plots, 3) effective sample sizes (Taboga 2017; Roy 2020), and 4) potential scale reduction factor (PSRF) (Gelman and Rubin 1992). Analyses and plots were done in R using the function *acf* for autocorrelation, and R package *coda* (Plummer *et al.* 2006) for effective sample sizes and potential scale reduction factor. These results were used to inform our decisions on burn-in and thinning for MCMC, as well as interpreting results of our evaluations of the methods under different simulated conditions.

ARGweaver

Convergence of likelihoods ARGweaver's *arg-sample* program outputs a *.stats* file containing several statistics for each MCMC iteration: log probability of the sampled ARG given the model ("prior", in Table S1), log probability of the data given the sampled ARG ("likelihood"), total log probability of the ARG and the data ("joint"), number of recombination events in the sampled ARG ("recombs"), the number of variant sites that cannot be explained by a single mutation under the sampled ARG ("noncompats"), total length of all branches summed across sites ("arglen") (Hubisz and Siepel 2020). We generated trace plots and calculated autocorrelation between consecutive samples using the likelihood per iteration (Figures S9 and S11). Following visual inspection of these plots, we chose a burn-in consisting of the first 200 samples in most simulations, except in simulations with 10 times higher mutation rate (Figure S9C,F) or sample sizes larger than 8 haplotypes (Figure S11B,C,E,F), where we chose a burn-in of 1200 samples since those chains took longer to converge. In both cases, we ran MCMC for 1000 iterations after burn-in. Based on autocorrelation plots (Figure S9, S11) and on effective sample sizes (Table S1), we thinned ARGweaver samples by recording every 10th MCMC iteration, thus retaining a total of 100 MCMC samples.

Results of the potential scale reduction factor suggested convergence of ARGweaver in simulations with mutation rate equal to recombination rate, with decreased recombination rate and with increased mutation rate (Table S1) - see section below on convergence of individual coalescence times.

Convergence of coalescence times For comparison with Relate, which does not output statistics for each iteration, we also analyse convergence of pairwise coalescence times in ARGweaver. To this end, we extract from each MCMC iteration the values of coalescence times between two pairs of samples at 100 sites equally spaced by 50 kb along the 5Mb simulated sequences. We use those 200 values for convergence diagnostics. Figure S12 shows trace plots of 10 of those sites, for one pair of samples. To evaluate convergence, we calculate potential scale reduction factor (PSRF) for each of the 200 coalescence times, and compare their mean, variance and range (Table S2) among different simulations. In Table S2 we also compare the number of coalescence times that have effective sample sizes lower than 100 (which is our MCMC sample size). These results also lead us to conclude that ARGweaver runs with mutation rate equal to recombination rate have converged. However, in contrast to the results on convergence for statistics recorded in the ARGweaver *stats* files (Table S1), the evaluation of convergence based on coalescence times does not support a conclusion of full convergence for the other simulated data sets. In particular, simulations with mutation to recombination rate ratio of 10 had a large number of coalescence times with effective sample size smaller than 100. The same was true for simulations with 16 and 32 haplotypes. The maximum values of PSRF in those simulations are also further from one, thus indicating a lack of convergence for some coalescence times.

Relate

Relate estimates branch lengths using an MCMC algorithm with built in burn-in (Speidel *et al.* (2019) Supplementary Note on Method details 4.2, p. 13). To obtain samples from the posterior distribution, the tree sequence estimated in this first step was used as a starting point. Therefore, we did not implement any extra burn-in to obtain samples from the posterior. Visual inspection of traces plots also suggested that additional burn-in was not necessary (Figure S13).

We evaluated Relate's MCMC convergence by running it 5 times for each sequence of 5Mb simulated under each set of parameters. We then extracted a subset of pairwise coalescence times to calculate the potential scale reduction factor and effective sample sizes as described above for ARGweaver. We extracted coalescence times for two pairs of samples at 100 equally spaced sites along the sequence (*i.e.* separated by 50kb). Table S3 shows these results, which indicate convergence of all Relate runs in all simulated datasets.

Tsdate prior grid

We ran tsdate with different prior grids, using the function tsdate.build_prior_grid(). The observation that dates inferred by tsdate seem to be bounded to a low maximum value still holds when changing prior grids to have more points (timepoints=100, Figure S14) or when manually specifying time slices with a maximum value of 12 (timepoints=np.geomspace(1e-5, 12, 50), Figure S15).

ARGweaver subtree sampling acceptance rates

As suggested by ARGweaver authors (Melissa Hubisz and Adam Siepel, personal communication), we have verified that acceptance rates of subtree sampling steps of ARGweaver are within a range that indicates good mixing of the chain, between 10% and 90% (Table

S4). All simulations except for the one with reduced recombination rate were within that range. For a visualization of the spread of the values of acceptance rate, Figure S16 shows the acceptance rates for subtree sampling steps of ARGweaver in one 5Mb region of each simulation.

Additional simulations results for ARGweaver

SMC and SMC' modes in ARGweaver

In all results shown in the main text, we simulated under the standard Hudson (1983) coalescent with recombination, and did inference in ARGweaver under SMC'. Here, we asked whether deviations observed in the posterior distribution of ARGweaver can be explained by differences between the models used for simulation and inference. For this, we simulate sequences in msprime under the SMC and SMC' models, and run ARGweaver inference using the same model used in the simulation. We simulated 8 haplotypes with mutation rate and recombination rate 2×10^{-8}. Results improve when simulating under SMC' and inferring under SMC' (Figures S17B, S18B). Surprisingly, simulating and inferring under SMC (Figures S17A, S18A) is not better than simulating under the full coalescent with recombination model and inferring under SMC (Figures 4, 5).

Intermediate values of mutation to recombination rate ratio

Rasmussen *et al.* (2014) mention in their Figure S5 that the quality of ARGweaver estimates generally improved in their simulations with increased mutation to recombination rates ratio (μ/ρ), but only up to $\mu/\rho = 4$. Motivated by this observation, we additionally ran simulations with values of μ/ρ in between the ones shown in the main text (μ/ρ=1 or μ/ρ=10), including μ/ρ=2 and 4. We summarize our results under these conditions in Table S5. We observed a similar pattern for these intermediate values of $\mu/\rho = 2, 4$ as we had observed from 1 to 10, *i.e.* point estimates improve with increased ratio (shown by lower MSE in Table S5), and calibration of the posterior distribution worsens with an increased ratio (show by higher KLD in Table S5).

Jukes-Cantor mutational model

In all results shown in the main text, we simulated mutations using an infinite sites model. ARGweaver, on the other hand, uses a Jukes and Cantor (1969) mutational model. Therefore, we hypothesize that differences in the mutational model between simulations and inference could explain deviations in the posterior distribution of ARGweaver, especially in simulations with increased mutation to recombination ratio (μ/ρ). We found that ARGweaver results with simulations under the Jukes and Cantor (1969) model are very similar to the results under the infinite sites model and follow the same pattern under increased μ/ρ (Table S5, Figures S20, S21).